# Expansion of the metazoan virosphere

# Expansion of the Metazoan Virosphere: Progress, Pitfalls, and Prospects

Darren J Obbard[1*]

[1]Institute of Evolutionary Biology, and Centre for Immunity, Infection and Evolution

The University of Edinburgh

Charlotte Auerbach Road

Edinburgh, EH9 3FL

*darren.obbard@ed.ac.uk

**Keywords:**

Virus; Metagenomics; RNAi; Host-range; Picornavirales;

**Abstract:**

Metagenomic sequencing has led to a recent and rapid expansion of the animal virome. It has uncovered a multitude of new virus lineages from under-sampled host groups, including many that break up long branches in the virus tree, and many that display unexpected genome sizes and structures. Although there are challenges to inferring the existence of a virus from a 'virus-like sequence', in the absence of an isolate the analysis of nucleic acid (including small RNAs) and sequence data can provide considerable confidence. As a consequence, this period of molecular natural history is helping to re-shape our views of deep virus evolution.

**Highlights:**

- Metagenomic discovery now contributes substantially to our view of virus evolution
- Serendipitous virus sequences from transcriptomes and genomes are under-utilised
- Small-RNA and strand-specific sequencing aid interpretation of viral metagenomes
- Barcode-switching and cryptic/unintended host material hamper robust host assignment
- Systematic host sampling is required to estimate the total number of animal viruses

**Explosive metagenomic growth**

It is 120 years since the word 'virus' was first applied specifically to a viral pathogen [1], but the number of known viruses is growing faster than ever (figure 1A; [2]). Much of this growth is through metagenomic discovery: the undirected large-scale sequencing of nucleic acids sampled from potential hosts or their environment [2-4]. Pioneered by studies of bacteriophage in the marine environment [5], recent years have witnessed an explosion in metagenomic sampling of the metazoan virosphere. This boom has focussed first on viruses likely to infect us and our livestock, particularly the virome of mammalian faeces [e.g. 6], on putative disease reservoirs such as bats [e.g. 7,8], and on arbovirus vectors [e.g. 9]. Subsequently, the focus has expanded to include neglected animal lineages, identifying hundreds of new RNA viruses in arthropods and other invertebrates [10-13], and recently in divergent and under-sampled chordates [14,15].

Compared to the isolation of new virus cultures, metagenomic discovery seems (relatively) cheap, easy, and (virtually) guaranteed—sequences often appear 'for free' when sequencing genomes and transcriptomes (Figure 1B-E) [10,16-18]. Nevertheless, there are clearly limitations to metagenomic discovery—especially for important applied questions such as "Where is the pandemic coming from?" [2]. With an isolate in hand we would have more than just a 'virus-like sequence': we could unambiguously confirm the host, be confident we hadn't been misled by a computational artefact, and study viral replication, host range and immunity [19-21]. However, our catalogue of the virosphere is in its infancy, and there are still great gains to be made from simple 'molecular natural history'. Fewer than 5 thousand viruses have received formal taxonomic recognition [22] and only around 15 thousand have even been named informally (Figure 1A). This is less comprehensive than the 17$^{th}$ century view of plant diversity, even in absolute terms [ca. 18 thousand species, 23], but few biologists today would claim the naturalists of subsequent centuries wasted their effort when making herbarium collections. And a modern evolutionary virologist can probably learn more from a virus genome than a 17$^{th}$ century botanist could from a dried specimen.

Metagenomic discovery has already had a huge impact on our knowledge of virus diversity. It has 'filled in' shallower parts of the tree, finding close relatives of iconic human pathogens, such as new influenzas in toads and eels [14]. It has also discovered new deep branches, such as clades of insect-infecting Partitiviruses [10,11] and Luteo/Sobemo-like viruses [10,24], and whole new families, such as the Chuviruses [25]. This in turn has led to renewed interest in inferring deep viral phylogenies [11,26], and has prompted proposals for large-scale updates of higher-level virus taxonomy [27]. More importantly, metagenomics now contributes to our thinking on virus evolution. It has provided a better perspective on host-association and host-switching [14,28,29], found familiar virus lineages with unexpected genome sizes and structures [11,25,30], and uncovered an unexpectedly dynamic history of 'modular' protein swapping [11,26]. Finally, merely having a PCR product from a metagenomic sample can provide an experimental route to the functional biology of an uncultured virus [31].

**Potential pitfalls**

The recent viral bonanza partly reflects advances in nucleic acid sequencing, a technology that has left Moore's Law—that computational power doubles every 2 years—far behind [32]. But sequencing is just one of the challenges to exploring the virosphere. The lack of a viable meta-barcoding sequence means that virus discovery often takes a full metagenomic approach, sequencing total (or virus-enriched) nucleic acid, and subsequently assigning sequences through inferred homology [e.g. 3,33,34]. This is challenging because high divergence means that only the most conserved sequences are recognisable (e.g. RNA virus polymerases), and even then, only at the protein level. Sensitive surveys therefore benefit from assembled contigs rather than raw reads (so that divergent genes are linked to recognisable ones) and protein rather than nucleic-acid similarity searches (because divergence is high). This can be done using off-the-shelf assemblers and search algorithms such as SPADes [35] or Trinity [36], and Diamond [37], but there is also a growing ecosystem of virus-specific metagenomic packages and pipelines available [34].

As with any field in rapid development, best practice is uncertain and fluid, and there are pitfalls for the unwary [3]. For example, although the assembly of virus (especially RNA virus) genomes is facilitated by their small size and largely unrepetitive nature, the high complexity of metagenomic pools tends to promote artefactual and chimeric contigs [4,38]. These can unite viral sequences with non-viral ones, especially high-copy-number host sequences such as those from mitochondria and ribosomes. Such 'wide' chimeras are partly mitigated by the use of paired-end and strand-specific reads, ensuring effective adaptor removal, and (when possible) removing host reads before assembly (although this can introduce problems if virus reads can cross-map to the host). Chimeric mis-assemblies among divergent viruses or viral segments are also possible, especially when they share near identical stretches of sequence, such as structural RNA motifs or terminal repeats. These are harder to diagnose, and may ultimately require PCR verification, but can often be flagged by comparison with close relatives (if available), unexpected local variation in read-depth, and comparison across metagenomic samples.

These challenges aside, discovering a 'virus-like sequence' remains easier than confirming its status as an infectious agent of the targeted host. First, even if a sequence is ultimately virus-derived in an evolutionary sense, its immediate origin may have been an Endogenous Viral Element (EVE) [39]. If expressed and/or 'domesticated' by the recipient genome, EVEs may be represented at high levels and retain open reading frames [39,40]. Equivalently, host sequences—especially transposable elements (TEs)—are often incorporated into large DNA viruses and can move freely between hosts and viruses [41], allowing these host sequences to be misclassified as viral in origin. Second, the host can be misassigned if samples contain multiple hosts, either naturally or through contamination. Although nucleic acid contamination is minimised by good laboratory practice, externally contaminated reagents [e.g. 42], nucleic acids involved in reagent production (e.g. reads from Murine leukaemia virus [43]) and library miss-assignment at the point of sequencing, can all be harder to identify and to exclude. In particular, 'barcode switching' (or 'hopping') in some Illumina platforms can misattribute reads among libraries at rates of up to 1% [44], and while this is reduced by incorporating barcodes in both adaptors ('dual indexing'), it is not always completely mitigated. Multi-host samples are often explicitly recognised as such, for example those from 'holobionts' such as anemones [45]. However, the multi-host nature of other samples is sometimes downplayed. For example, faecal samples are often dominated by viruses infecting the gut microbiota and/or organisms in the host's diet [46,47], but virus-like sequences are sometimes reported (at least in the headline) as if they were viruses of the faecal donor itself. And, if nucleic acids or virions are prepared from whole host individuals, viruses in faecal matter and viral infections of parasites (notably nematodes, platyhelminthes, and microscopic arthropods) and pathogens (fungi, trypanosomatids, apicomplexans, amoebae, and many others) will also be represented among the sequences. Pre-screening of samples for specific parasites by PCR [such as nematodes, e.g. 13]—can mitigate against this, as can tissue dissection [14] (although at the potential risk of biasing discovery toward viruses with a strong tissue tropisms). However, the potential for viral infections of eukaryotic parasites within the metazoan host means that even dissected tissue may be cryptically multi-host. For example, the only dimarhabdovirus recorded from a plant sample derives from RNA contaminated with thrips [16].

**Going beyond 'virus-like sequences'**

Such pitfalls make some authors (justifiably) hesitant to proclaim a new virus from metagenomic sequencing alone, and many instead choose to report 'virus-like sequences'—providing an implicit *caveat emptor*. But even in the absence of an isolate, sequence data and nucleic acid analysis can be used to support the existence of a free/replicating virus. First, the nature and quantity of the nucleic acid provides useful clues. Endogenous DNA copies can be identified by a comparison of PCR and RT-PCR (or direct DNA and RNA sequencing) [10,11,13,48]. For example, functional DNA viruses must express their proteins, so that the absence of viral mRNAs argues against active replication. Active replication also affects strand-bias in RNA viruses, so that strand-specific PCR [49] or RNA sequencing can identify the negative-sense replication intermediates of positive-

sense single-stranded (+ss) RNA viruses, and quantitative analyses can detect the presence of coding products from -ssRNA and dsRNA viruses [48]. And, for both DNA and RNA viruses, contaminating sequences are likely to be at relatively low titre, whereas the copy-number of inherited EVEs will match the host genome. This means that high copy-number itself provides an argument in favour of viral status [11], especially when viruses can contribute more than 10%, and sometimes in excess of 50%, of total (non-ribosomal) RNA in some species [50].

Second, contigs that encode complete viral genomes with intact open reading frames are more consistent with functional viruses than with EVEs. And, although whole viruses can be (retro-)copied into a host's genome, there is rarely selective pressure to maintain the virus genome intact—resulting in segregating frameshift and nonsense mutations. Even expressed and functional (i.e. 'domesticated') EVEs generally only provide the host with one or two beneficial sequences [39,40]. Complete or near-complete virus genomes can also rule out the misattribution of host TEs as viral sequences, as the larger DNA viruses that carry TEs are unlikely to be represented by their TE sequences alone. Third, the distribution of virus-like sequences across metagenomic pools and host individuals (e.g. surveyed by PCR) can help to confirm a genuine viral origin [10,13,50]. Presence/absence patterns can help to weed out EVEs, as—unless it is very recent in origin—an EVE insertion is likely to be present in all host genomes, whereas virus prevalence is likely be below 100% and variable among populations and over time [10,13]. The co-occurrence of virus-like and other sequences across host individuals can be used to correctly infer hosts, as viruses that infect a contaminating microparasite will co-occur with it. Similarly, patterns of co-occurrence can also help to identify missing parts of the viral genome, such as fragments of incompletely assembled genomes and components of segmented viruses that are not recognisable using sequence similarity [10,50].

Finally, perhaps the ultimate evidence of infection is recognition by the host antiviral immune system [10,51,52]. In vertebrates, the presence antibodies can be used to corroborate infection [51]. In nematodes and arthropods, the distinctive small RNAs (viRNAs) generated from viral genomes by antiviral RNA-interference (RNAi) [53,54] can be used in a similar way. Because Dicer-mediated viRNA biogenesis targets dsRNA such as replication intermediates, viRNAs can demonstrate both an antiviral response and viral replication. Importantly, viRNAs usually have a tight and characteristic length distribution (e.g. 20nt in Lepidoptera, 21nt in *Drosophila*, 22nt in C. elegans) [53,54] and a 3' 2-O-methyl group, making them distinguishable from degradation products. Their size distribution and base composition also distinguish them from TE- and EVE-derived piwi-associated RNAs [13,40,54]. Notably, and unlike antibodies, viRNAs can identified directly from the metagenomic discovery RNA pool, allowing confirmation concomitantly with metagenomic discovery [10,52].

**How many animal viruses are there, and what are they doing?**

Our expanded view of the animal virosphere has already started to answer old questions and provoke new ones, but these two stand out. What prospect is there of answering them? Given any definition of 'different virus' [4,21], whether based on an operational taxonomic unit or a functional biological definition, virus lineages are countable. Sampling of nine virus families to near-saturation from one bat species in Bangladesh identified 55 different viruses and implied an estimate of 320 thousand viruses infecting mammals [55]. However, if a substantial proportion of these viruses were either multi-host or represented recent spillover from other hosts (i.e without onward transmission) the estimate would be very different, and the estimate might also be biased by the particular choice of virus families and geographic region. A more confident estimate could be made from unbiased metagenomic samples of the joint distribution of prevalence across host and virus lineages, sampled across their geographic range. For example, near-saturation sampling of multiple taxonomic groups within a single ecosystem-type across a geographic region would not only allow the virus diversity within host lineages to be assessed, but would also allow an assessment of host range and—from sequence analysis—the timescale of host switching.

Such metagenomic surveys may soon be possible for a few carefully-considered host groups,

but they would still miss those virus sequences that we cannot see because they lack detectable homology with known viruses, the so-called viral 'dark matter' [56,57]. Many of these 'dark matter' sequences, perhaps the majority, are likely to represent the poorly-conserved regions of otherwise recognisable viruses [57,58]. However, some completely new and/or highly divergent virus lineages, which cannot be detected using the *de facto* default choice of search tools and parameters, probably remain to be discovered. Indeed, several recently-identified viruses were initially detected from contig size and nucleic acid abundance, and only subsequently attributed as viral using higher-sensitivity similarity searches [11]. In the future, more powerful search tools, such as those based on protein profile-profile comparisons [57,59], and a search for deeper homologies, such as those provided by protein structure [60], may prove useful. Where no remaining protein similarity exists, a complementary approach to is to consider sequences that are flagged as potentially viral in origin by the antiviral RNAi immune response of plants, insects, and fungi. Webster *et al* [10] proposed around 60 such viRNA-based 'candidate virus' contigs based on metagenomic sequencing from *Drosophila*, and approximately half of these have since been identified as fragments from known virus lineages by the subsequent discovery of related viruses, or by an analysis of co-occurrence across samples [e.g. 50]. This leaves open the possibility that some do represent genuinely new viruses (e.g. Figure 1E), but the ultimate confirmation of genuinely novel virus lineages probably represents a case in which viral isolates are unequivocally necessary.

What are these viruses doing to their hosts? It is almost axiomatic that viruses are parasites, but micro-organisms are often mutualist or commensal, and although viruses necessarily use host resources, their impact on host fitness may be negligible and/or outweighed by provision of some unknown benefit [61]. It might initially seem that elucidating the fitness consequences of infection must also require isolates for experimentation. However, experimental studies are rarely useful for inferring real-world fitness. First, most studies measure traits such as survival or reproduction in place of fitness. This can misinterpret life-history tradeoffs, such as mistaking a host response to mitigate cost (e.g. terminal investment) for a virus-derived benefit (increased early-life reproduction). Second, such studies tend to be under-powered: an absence of detectable harm does not imply costs are absent, only that they are small. But at what point is a cost so small that the virus is effectively commensal? The ultimate arbiter of costliness must be natural selection: if the presence of the virus selects for host resistance, then the virus imposes a net fitness cost, by definition. A resistance mutation is expected to spread if its fitness benefit substantially exceeds the impact of genetic drift (i.e. $N_e s >> 1$ where $N_e$ is effective population size and $s$ is the selective benefit). Very conservatively, an infection cost of 0.1% in *Drosophila* (or many other small invertebrates with large effective population size) would select strongly for host resistance. However, this cost is probably an order of magnitude too small to measure experimentally in a multicellular organism [62], meaning that it is effectively impossible to experimentally distinguish between a low-cost virus and a commensal one. Far from requiring more isolates, the best solution to understanding fitness consequences of infection could also be a metagenomic one, by adding metagenomic screens to fitness studies of animals in the wild [e.g. 63].

**Acknowledgements**

I thank Alistair Greaves for preparing the phylogeny of Picornavirales in Figure 1B and Ben Longdon, Fergal Waldron, Mang Shi, David Karlin and two anonymous reviewers for comments. I apologise to the many authors whose work could not be cited due to restrictions on space and publication timeframe.

**Funding**

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Bibliography**


1. Bos L: **Beijerinck's work on tobacco mosaic virus: historical context and legacy**. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 1999, **354**:675-685.
2. Greninger AL: **A decade of RNA virus metagenomics is (not) enough**. *Virus Research* 2018, **244**:218-229.

** A recent, comprehensive, and highly entertaining perspective on virus metagenomics. This is the review I wanted to write.

3. Rose R, Constantinides B, Tapinos A, Robertson DL, Prosperi M: **Challenges in the analysis of viral metagenomes**. *Virus Evolution* 2016, **2**:vew022-vew022.
4. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, Davison AJ, Delwart E, Gorbalenya AE, Harrach B, et al.: **Virus taxonomy in the age of metagenomics**. *Nature Reviews Microbiology* 2017, **15**:161.

** A thorough exploration of the impact on metagenomic discovery on our understanding of virus diversity

5. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, Azam F, Rohwer F: **Genomic analysis of uncultured marine viral communities**. *Proceedings of the National Academy of Sciences* 2002, **99**:14250-14255.
6. Williams SH, Che XY, Garcia JA, Klena JD, Lee B, Muller D, Ulrich W, Corrigan RM, Nichol S, Jain K, et al.: **Viral Diversity of House Mice in New York City**. *Mbio* 2018, **9**:17.
7. Berto A, Anh PH, Carrique-Mas JJ, Simmonds P, Van Cuong N, Tue NT, Van Dung N, Woolhouse ME, Smith I, Marsh GA, et al.: **Detection of potentially novel paramyxovirus and coronavirus viral RNA in bats and rats in the Mekong Delta region of southern Viet Nam**. *Zoonoses and Public Health* 2018, **65**:30-42.
8. Zheng XY, Qiu M, Guan WJ, Li JM, Chen SW, Cheng MJ, Huo ST, Chen Z, Wu Y, Jiang LN, et al.: **Viral metagenomics of six bat species in close contact with humans in southern China**. *Archives of Virology* 2018, **163**:73-88.
9. Tokarz R, Sameroff S, Tagliafierro T, Jain K, Williams SH, Cucura DM, Rochlin I, Monzon J, Carpi G, Tufts D, et al.: **Identification of Novel Viruses in Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis Ticks**. *Msphere* 2018, **3**.
10. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, Brouqui JM, Bayne EH, Longdon B, Buck AH, et al.: **The Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster**. *Plos Biology* 2015, **13**:33.
11. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, Qin XC, Li J, Cao JP, Eden JS, et al.: **Redefining the invertebrate RNA virosphere**. *Nature* 2016, **540**:539-+.

** The largest single report of new animal virus diversity to date, providing an exceptional illustration of the power of viral metagenomics approaches in animals

12. Roberts JMK, Anderson DL, Durr PA: **Metagenomic analysis of Varroa-free Australian honey bees (Apis mellifera) shows a diverse Picornavirales virome**. *Journal of General Virology* 2018.
13. Waldron FM, Stone GN, Obbard DJ: **Metagenomic sequencing suggests a diversity of RNA interference-like responses to viruses across multicellular eukaryotes**. *PLOS Genetics* 2018, **In Press**.
14. Shi M, Lin XD, Chen X, Tian JH, Chen LJ, Li K, Wang W, Eden JS, Shen JJ, Liu L, et al.: **The evolutionary history of vertebrate RNA viruses**. *Nature* 2018, **556**:197-+.

* Reports more than 200 new viruses from chordates, demonstrating the importance of targeted taxon sampling

15. Geoghegan JL, Pirotta V, Harvey E, Smith A, Buchmann JP, Ostrowski M, Eden J, Harcourt R, Holmes EC: **Virological Sampling of Inaccessible Wildlife with Drones**. 2018:Preprints 2018, 2018050184.
16. Longdon B, Murray GGR, Palmer WJ, Day JP, Parker DJ, Welch JJ, Obbard DJ, Jiggins FM: **The evolution, diversity, and host associations of rhabdoviruses**. *Virus Evolution* 2015, **1**:12.
17. François S, Filloux D, Roumagnac P, Bigot D, Gayral P, Martin DP, Froissart R, Ogliastro M: **Discovery of parvovirus-related sequences in an unexpected broad range of animals**. *Scientific Reports* 2016, **6**:30880.

18. Kapun M, Barron Aduriz MG, Staubach F, Vieira J, Obbard D, Goubert C, Rota Stabelli O, Kankare M, Haudry A, Wiberg RAW, et al.: **Genomic analysis of European Drosophila populations reveals longitudinal structure and continent-wide selection**. *bioRxiv* 2018.
19. Ladner JT, Beitzel B, Chain PSG, Davenport MG, Donaldson E, Frieman M, Kugelman J, Kuhn JH, O'Rear J, Sabeti PC, et al.: **Standards for Sequencing Viral Genomes in the Era of High-Throughput Sequencing**. *mBio* 2014, **5**.
20. Murphy FA: **Chapter Five - Historical Perspective: What Constitutes Discovery (of a New Virus)?** In *Advances in Virus Research*. Edited by Kielian M, Maramorosch K, Mettenleiter TC: Academic Press; 2016:197-220. vol 95.]
21. van Regenmortel MH: **Classes, taxa and categories in hierarchical virus classification: a review of current debates on definitions and names of virus species**. *Bionomina* 2016, **10**:1-21.
22. King AMQ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, Gorbalenya AE, Harrach B, Harrison RL, Junglen S, Knowles NJ, et al.: **Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018)**. *Archives of Virology* 2018.
23. Ray J: *Methodus Plantarum Nova*: The Ray Society; 2014.
24. Tokarz R, Williams SH, Sameroff S, Sanchez Leon M, Jain K, Lipkin WI: **Virome Analysis of Amblyomma americanum, Dermacentor variabilis, and Ixodes scapularis Ticks Reveals Novel Highly Divergent Vertebrate and Invertebrate Viruses**. *Journal of Virology* 2014, **88**:11480-11492.
25. Li CX, Shi M, Tian JH, Lin XD, Kang YJ, Chen LJ, Qin XC, Xu JG, Holmes EC, Zhang YZ: **Unprecedented genomic diversity of RNA viruses in arthropods reveals the ancestry of negative-sense RNA viruses**. *Elife* 2015, **4**.
26. Koonin EV, Dolja VV, Krupovic M: **Origins and evolution of viruses of eukaryotes: The ultimate modularity**. *Virology* 2015, **479-480**:2-25.

***** Demonstrates what metagenomic discovery might tell us about mechanisms of evolution

27. Aiewsakun P, Simmonds P: **The genomic underpinnings of eukaryotic virus taxonomy: creating a sequence-based framework for family-level virus classification**. *Microbiome* 2018, **6**:38.

***** Elucidates the likely impact of new-found viral diversity on higher-level virus taxonomy

28. Geoghegan JL, Duchêne S, Holmes EC: **Comparative analysis estimates the relative frequencies of co-divergence and cross-species transmission within viral families**. *PLOS Pathogens* 2017, **13**:e1006215.
29. Dolja VV, Koonin EV: **Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer**. *Virus research* 2018, **244**:36-52.
30. Shi M, Lin XD, Vasilakis N, Tian JH, Li CX, Chen LJ, Eastwood G, Diao XN, Chen MH, Chen X, et al.: **Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses**. *Journal of Virology* 2016, **90**:659-669.
31. van Mierlo JT, Overheul GJ, Obadia B, van Cleef KWR, Webster CL, Saleh MC, Obbard DJ, van Rij RP: **Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi**. *Plos Pathogens* 2014, **10**:13.
32. Wetterstrand K: **DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)** https://www.genome.gov/sequencingcostsdata. 14th May 2018
33. Paez-Espino D, Pavlopoulos GA, Ivanova NN, Kyrpides NC: **Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data**. *Nature Protocols* 2017, **12**:1673.
34. Nooij S, Schmitz D, Vennema H, Kroneman A, Koopmans MPG: **Overview of Virus Metagenomic Classification Methods and Their Biological Applications**. *Frontiers in Microbiology* 2018, **9**.

***** A comprehensive and systematic overview of software intended for the analysis of viral metagenomic data

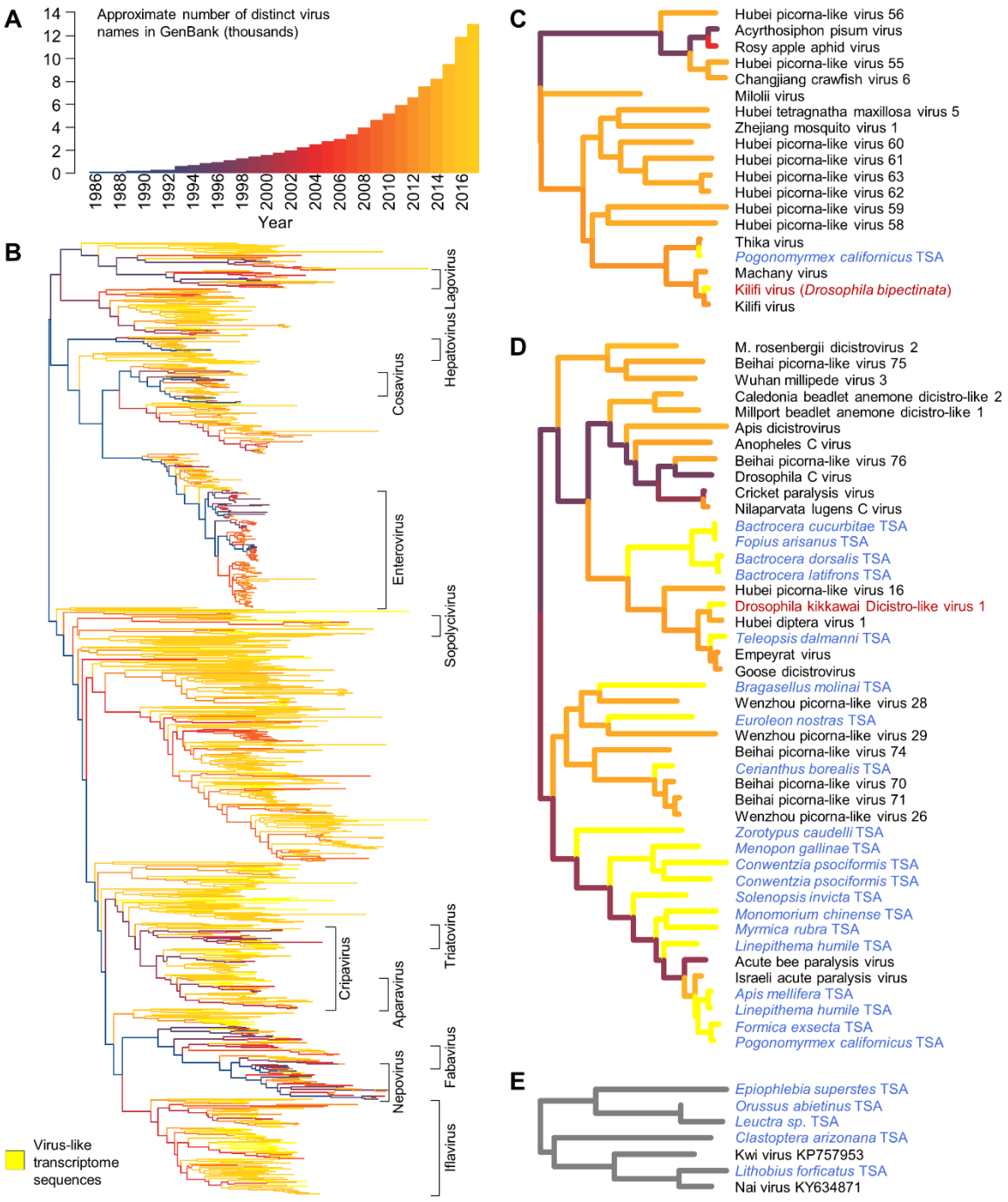
35. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD: **SPAdes: a new genome**

**assembly algorithm and its applications to single-cell sequencing**. *Journal of computational biology* 2012, **19**:455-477.
36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q: **Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data**. *Nature biotechnology* 2011, **29**:644.
37. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND**. *Nature Methods* 2014, **12**:59.
38. Tithi SS, Aylward FO, Jensen RV, Zhang L: **FastViromeExplorer: a pipeline for virus and phage identification and abundance profiling in metagenomics data**. *PeerJ* 2018, **6**:e4227.
39. Katzourakis A, Gifford RJ: **Endogenous Viral Elements in Animal Genomes**. *PLOS Genetics* 2010, **6**:e1001191.
40. Palatini U, Miesen P, Carballar-Lejarazu R, Ometto L, Rizzo E, Tu Z, Rij RP, Bonizzoni M: **Comparative genomics shows that viral integrations are abundant and express piRNAs in the arboviral vectors Aedes aegypti and Aedes albopictus**. *BMC genomics* 2017, **18**:512.
41. Gilbert C, Peccoud J, Chateigner A, Moumen B, Cordaux R, Herniou EA: **Continuous Influx of Genetic Material from Host to Virus Populations**. *PLOS Genetics* 2016, **12**:e1005838.
42. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, Aronsohn A, Hackett J, Delwart EL, Chiu CY: **The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns**. *Journal of Virology* 2013, **87**:11966-11977.
43. Ge X, Li Y, Yang X, Zhang H, Zhou P, Zhang Y, Shi Z: **Metagenomic Analysis of Viruses from Bat Fecal Samples Reveals Many Novel Viruses in Insectivorous Bats in China**. *Journal of Virology* 2012, **86**:4620-4630.
44. Kircher M, Sawyer S, Meyer M: **Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform**. *Nucleic Acids Research* 2012, **40**:e3-e3.
45. Brüwer JD, Voolstra CR: **First insight into the viral community of the cnidarian model metaorganism Aiptasia using RNA-Seq data**. *PeerJ* 2018, **6**:e4449.
46. Zhang T, Breitbart M, Lee WH, Run JQ, Wei CL, Soh SWL, Hibberd ML, Liu ET, Rohwer F, Ruan YJ: **RNA viral community in human feces: Prevalence of plant pathogenic viruses**. *Plos Biology* 2006, **4**:108-118.
47. Li LL, Victoria JG, Wang CL, Jones M, Fellers GM, Kunz TH, Delwart E: **Bat Guano Virome: Predominance of Dietary Viruses from Insects and Plants plus Novel Mammalian Viruses**. *Journal of Virology* 2010, **84**:6955-6965.
48. Medd NC, Fellous S, Waldron FM, Xuéreb A, Nakai M, Cross JV, Obbard DJ: **The virome of Drosophila suzukii, an invasive pest of soft fruit**. *Virus Evolution* 2018, **4**:vey009-vey009.
49. Plaskon NE, Adelman ZN, Myles KM: **Accurate Strand-Specific Quantification of Viral RNA**. *PLOS ONE* 2009, **4**:e7468.
50. Shi M, White VL, Schlub T, Eden J-S, Hoffmann AA, Holmes EC: **No detectable effect of Wolbachia wMel on the prevalence and abundance of the RNA virome of *Drosophila melanogaster***. *Proceedings of the Royal Society B-Biological Sciences* 2018, **In Press**.
51. Burbelo PD, Ching KH, Esper F, Iadarola MJ, Delwart E, Lipkin WI, Kapoor A: **Serological Studies Confirm the Novel Astrovirus HMOAstV-C as a Highly Prevalent Human Infectious Agent**. *PLOS ONE* 2011, **6**:e22576.
52. Aguiar E, Olmo RP, Paro S, Ferreira FV, de Faria IJD, Todjro YMH, Lobo FP, Kroon EG, Meignin C, Gatherer D, et al.: **Sequence-independent characterization of viruses based on the pattern of viral small RNAs produced by the host**. *Nucleic Acids Research* 2015, **43**:6191-6206.
53. Félix M-A, Ashe A, Piffaretti J, Wu G, Nuez I, Bélicard T, Jiang Y, Zhao G, Franz CJ, Goldstein LD: **Natural and experimental infection of Caenorhabditis nematodes by novel viruses related to nodaviruses**. *PLoS biology* 2011, **9**:e1000586.
54. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, Sharma PP, Cordaux R, Gilbert C, Giraud I: **Pan-arthropod analysis reveals somatic piRNAs as an ancestral defence against transposable elements**. *Nature ecology & evolution* 2018, **2**:174.
55. Anthony SJ, Epstein JH, Murray KA, Navarrete-Macias I, Zambrana-Torrelio CM, Solovyov A, Ojeda-Flores R, Arrigo NC, Islam A, Ali Khan S,

et al.: **A Strategy To Estimate Unknown Viral Diversity in Mammals**. *mBio* 2013, **4**.
56. Krishnamurthy SR, Wang D: **Origins and challenges of viral dark matter**. *Virus research* 2017, **239**:136-142.
57. Bzhalava Z, Hultin E, Dillner J: **Extension of the viral ecology in humans using viral profile hidden Markov models**. *PLOS ONE* 2018, **13**:e0190938.
58. François S, Filloux D, Frayssinet M, Roumagnac P, Martin DP, Ogliastro M, Froissart R: **Increase in taxonomic assignment efficiency of viral reads in metagenomic studies**. *Virus research* 2018, **244**:230-234.
59. Kuchibhatla DB, Sherman WA, Chung BYW, Cook S, Schneider G, Eisenhaber B, Karlin DG: **Powerful Sequence Similarity Search Methods and In-Depth Manual Analyses Can Identify Remote Homologs in Many Apparently "Orphan" Viral Proteins**. *Journal of Virology* 2014, **88**:10-20.
60. Yutin N, Bäckström D, Ettema TJG, Krupovic M, Koonin EV: **Vast diversity of prokaryotic virus genomes encoding double jelly-roll major capsid proteins uncovered by genomic and metagenomic sequence analysis**. *Virology Journal* 2018, **15**:67.
61. Roossinck MJ, Bazán ER: **Symbiosis: Viruses as Intimate Partners**. *Annual Review of Virology* 2017, **4**:123-139.
62. Gallet R, Cooper TF, Elena SF, Lenormand T: **Measuring selection coefficients below 10− 3: method, questions, and prospects**. *Genetics* 2012, **190**:175-186.
63. Knowles SC, Fenton A, Pedersen AB: **Epidemiology and fitness effects of wood mouse herpesvirus in a natural host population**. *Journal of General Virology* 2012, **93**:2447-2456.
64. Ma S, Avanesov AS, Porter E, Lee BC, Mariotti M, Zemskaya N, Guigo R, Moskalev AA, Gladyshev VN: **Comparative transcriptomics across 14 Drosophila species reveals signatures of longevity**. *Aging cell* 2018:e12740.

**Figure 1**

**Panel A**: The number of distinct names for viruses (excluding phage) in the GenBank nucleotide database, by year (colours provide a scale for Panels B-D). Counts were obtained by finding the record creation date and GenBank 'species' (collapsing strain identifiers) for each of 2.6 million virus sequences. Exclusion of unrecognised species names and the merging of divergent strains are likely to make this an underestimate. **Panel B:** Midpoint-rooted maximum likelihood phylogeny of picorna-like viruses and caliciviruses, inferred from approximately 250 amino acids of the polymerase. Branches are coloured by the year in which the lineage was first recorded in GenBank (colours provided by panel A). Approximately 8000 picorna-like polymerase sequences from the NCBI non-redundant protein (nr) and transcriptome shotgun assembly (tsa_nt) databases were identified by blastp and tblastn. These were collapsed into 1140 clusters at a threshold of 96% identity, with one representative of each cluster used to infer the tree. Around 10% of the represented picorna-like lineages are known only as unannotated virus-like sequences from transcriptomes (pale yellow; viruses from transcriptome datasets are treated as unpublished and given a more recent date). Note that the short conserved-sequence length leads to poor resolution and fails to recover some named genera, and that similarity criteria for inclusion means that some picornavirus groups were excluded. **Panels C and D:** To illustrate with ease with which new virus-like sequences can be found in public datasets, I obtained the most recently deposited *Drosophila* RNAseq dataset (PRJNA414017 [64]), performed a *de novo* assembly using Trinity [36], and identified virus-like sequences using Diamond [37]. I found complete genomes for two picorna-like viruses (red labels; MH320557 and MH320558): a divergent sequence of Kilifi virus from *D. bipectinata* (previously known from *D. melanogaster*) and a Dicistro-like virus from *D. kikkawai*, related to Hubei diptera virus 1 [11]. Maximum-likelihood phylogenies for these two sequences were inferred from around 700 amino acids of the polymerase, mid-point rooted, and coloured as in panel B. These trees illustrate the dominance of recent discoveries, including the many virus-like sequences in transcriptome assemblies (blue taxon labels). They also illustrate the potential confusion introduced by naming faecal-sample viruses after the faecal donor (all close relatives of Goose Dicistrovirus infect invertebrates). **Panel E:** Phylogeny of two putative 'dark matter' viruses from *Drosophila,* including related transcriptome sequences. These putative viruses each comprise four 1.5Kb segments encoding a single long open reading frame (the most conserved of which was used for phylogenetic inference), but they lack detectable homology with any known virus lineage and were inferred to be viral on the basis of viRNA profiles and co-occurrence across samples [10]. Data associated with this figure are available from via FigShare https://dx.doi.org/10.6084/m9.figshare.6272066 .


A
Approximate number of distinct virus names in GenBank (thousands)
14
12
10
8
6
4
2
0
1986
1988
1990
1992
1994
1996
1998
2000
2002
2004
2006
2008
2010
2012
2014
2016
Year
B
Lagovirus
Hepatovirus
Cosavirus
Enterovirus
Sopolycivirus
Triatovirus
Cripavirus
Aparavirus
Fabavirus
Nepovirus
Iflavirus
Virus-like transcriptome sequences
C
Hubei picorna-like virus 56
Acyrthosiphon pisum virus
Rosy apple aphid virus
Hubei picorna-like virus 55
Changjiang crawfish virus 6
Milolii virus
Hubei tetragnatha maxillosa virus 5
Zhejiang mosquito virus 1
Hubei picorna-like virus 60
Hubei picorna-like virus 61
Hubei picorna-like virus 63
Hubei picorna-like virus 62
Hubei picorna-like virus 59
Hubei picorna-like virus 58
Thika virus
Pogonomyrmex californicus TSA
Machany virus
Kilifi virus (Drosophila bipectinata)
Kilifi virus
D
M. rosenbergii dicistrovirus 2
Beihai picorna-like virus 75
Wuhan millipede virus 3
Caledonia beadlet anemone dicistro-like 2
Millport beadlet anemone dicistro-like 1
Apis dicistrovirus
Anopheles C virus
Beihai picorna-like virus 76
Drosophila C virus
Cricket paralysis virus
Nilaparvata lugens C virus
Bactrocera cucurbitae TSA
Fopius arisanus TSA
Bactrocera dorsalis TSA
Bactrocera latifrons TSA
Hubei picorna-like virus 16
Drosophila kikkawai Dicistro-like virus 1
Hubei diptera virus 1
Teleopsis dalmanni TSA
Empeyrat virus
Goose dicistrovirus
Bragasellus molinai TSA
Wenzhou picorna-like virus 28
Euroleon nostras TSA
Wenzhou picorna-like virus 29
Beihai picorna-like virus 74
Cerianthus borealis TSA
Beihai picorna-like virus 70
Beihai picorna-like virus 71
Wenzhou picorna-like virus 26
Zorotypus caudelli TSA
Menopon gallinae TSA
Conwentzia psociformis TSA
Conwentzia psociformis TSA
Solenopsis invicta TSA
Monomorium chinense TSA
Myrmica rubra TSA
Linepithema humile TSA
Acute bee paralysis virus
Israeli acute paralysis virus
Apis mellifera TSA
Linepithema humile TSA
Formica exsecta TSA
Pogonomyrmex californicus TSA
E
Epiophlebia superstes TSA
Orussus abietinus TSA
Leuctra sp. TSA
Clastoptera arizonana TSA
Kwi virus KP757953
Lithobius forficatus TSA
Nai virus KY634871